数字图书馆学者库构建方式研究*

郑昂 曾建勋
中国科学技术信息研究所 北京 100038

摘 要:[目的/意义]从数字图书馆资源利用与整理角度出发设计学者数据识别与学者数据库的构建方式,帮助提升数字图书馆资源建设效率与特色服务。[方法/过程]从学者遴选与收录来源、学者描述内容及其框架、学者库构建与学者库应用方式四个方面调研国内外学者库研究及实践情况。通过分析学者特征属性,研究学者数据结构化表达方式,提出基于数字图书馆的学者库构建流程和总体框架。[结果/结论]提出学者库构建与应用齐头并进的推进策略,强调学者库要融入科研管理过程,发动学者参与建设,增加展示与宣传效果,与人才识别相结合,服务于团队和专题资源建设;与知识管理相结合,兼顾学者存档与学者画像功能,拓展精准服务功能。

关键词:学者库 数字图书馆 机构知识库 学者识别
分类号:G250.74
DOI:10.13266/j.issn.0252-3116.2020.05.014

学者库以学者为资源组织对象开展学术特征信息描述,是存储、检索、利用和发现学者科研产出的数据系统。学者库不仅对学者学术特征、属性和学术成果进行著录识别,而且对学者学术关系、学术生涯和学术轨迹进行描述、链接;其形成的学术资源集,不仅是构建机构知识库的基础单元,也是评价学者学术绩效的基本元素,还是展示机构实力和学者风采的基本素材[1]。因此,学者库构建不仅是数字图书馆与科研平台特色资源建设的重要课题,还成为图书馆和科研组织精准服务于科研评价和科研人员的重要措施。

近年来,数据库商、高校和科研机构及部分科研项目资助机构都基于自建的数字图书馆资源和平台开展了学者库构建的实践探索。其中,高校和科研机构及科研项目资助常以满足自身的需求为导向,构建中采用人工方式,或借助数字图书馆技术与平台提升学者库的自动化水平;数据库商的学者库构建则注重满足各类用户的应用需求,全面覆盖各学科、机构的学者,推进构建流程的自动化,是当前研究与实践的重点。然而以商业数据库或知识机构库资源为基础数据,仅侧重于学术产出的集成与计量,存在无法全面揭示学者特征和无法全面涵盖学者学术成果的现象。为此,本文拟在完善学者库元数据体系的基础上,优化基于数字图书馆的学者库总体框架和构建流程设计,采用多源数据整合的方式优化学者库构建的基础资源,并提出学者库构建与应用推进策略。

1 国内外学者库建设现状

学者库的建设主要涉及学者科研活动、交流行为、学术关系、产出成果的揭示,以及学者评价与展示、学者识别与服务等方面。本文对国内外学者库收录范围、学者遴选方式进行分析,对学者描述系统和体系进行调研,对学者库构建方式及应用现状进行梳理,分析学者库构建的必要步骤。其中,学者遴选方式和收录来源影响学者库构建效果,学者特征描述是构建学者库的关键环节和有效应用的前提。

1.1 学者遴选与收录来源

不同的构建目标使学者库拥有不同的学者遴选范围和资源获取方式。商业数据库和学术搜索引擎根据一定的筛选条件,选择具有科研成果的学者为目标学者建立学者库。AMiner以人工智能等领域专家为目标遴选范围,将相关领域的论文进行集成整合,通过大规模的计算得出目标学者。百度学术为具有一定发文量与被引量的学者自动聚合学术成果,其他学者也能通过认领成果构建自己的主页,目前共生成400多万

* 本文系国家社会科学基金项目"多源异构数据融合的图书馆用户画像研究"(项目编号:18BTQ031)研究成果之一。
作者简介:郑昂(ORCID:0000-0001-8326-3858),硕士研究生,E-mail:zhengang2017@istic.ac.cn;曾建勋(ORCID:0000-0002-0432-961),信息资源中心主任,研究馆员,博士生导师。
收稿日期:2019-06-17 修回日期:2019-09-02 本文起止页码:133-140 本文责任编辑:杜杏叶

个学者主页。通过自动聚合学者信息,商业数据库和学术搜索引擎构建了大量的学者页面,但学者认领学术成果和页面数量较少,如中国知网学者库汇集了1 200万学者,但仅有 10 万人认领成果信息[2]。高校、科研机构所构建的学者库则以本单位学者为学者遴选范围,如西安交通大学 XJTU Academic Hub 规定提交者身份限定于本校教师、科研人员、在读研究生、本科生及本校其他教工[3]。

学术成果的收录范围影响学者库构建的效果。数据库商常以其收录的数据为基础,如中国知网学者库以 CNKI 中文期刊全文数据库为基础。这种方式受限于其收录资源的范围,无法全面揭示学者的学术成果,也难以涵盖学术成果之外的学者信息,而集成整合多源数据能够获得更丰富、完整的学者成果。百度学术发挥数字图书馆分布式资源与运行技术优势,通过与内容供应商合作获取题录数据,采用 AI-PMH 协议等的元数据收割技术对开放资源进行收割,并通过搜索引擎爬取数据,集合学者中外学术成果。对于高校和科研机构来说,其常以购买和自建的数字学术资源为基础进行构建,如清华大学、西安交通大学等高校以 WOS、EI、Nature、Science 等数据库为收录范围,并与机构知识库的科研成果资源相结合[3],同时辅以机构学者提交的个人信息,这种方式在初始建设环节常常可以收到不错效果,但后续的维护更新难以保证信息的时效性和准确性。

1.2 学者描述内容及其框架

学者库的构建需要对学者特征、成果和关系进行组织和描述,以此实现学者库的展示和应用。数据商、科研机构和学者唯一标识符系统通过对数据库资源整合、网络爬取、科研成果登记等不同方式对学者数据进行集成,描述内容和效果具有差异:ResearcherID、ORCID 等唯一标识符面向全球学者,能够最大范围地展示学者引文、合著等学术合作关系[4]。数据商、高校与科研机构构建的学者库对学者发文、被引等描述项揭示较为充分、及时,主要集成了学术经历、发文量分布、学科主题、合作者等特征信息等[5],基于数据库和知识库的资源优势,提供全文或链接。高校与机构学者库通过本单位获取职务、职称、荣誉等较为全面的学者基本信息。

当前大部分的学者库从数据库抽取学者机构、合作者等信息,对数据库中学者相关的文献元数据进行动态计量,但各科研实体间的关系揭示不够充分,没有从语义层面对学者信息进行推理、补充。而一些知识发现服务搜索系统为提高学者语义信息的抽取与描述,构建可存储、可运算的学者描述框架,实现学者及相关科研实体、关系的表达,可以成为优化数字图书馆学者描述与揭示方式的参考。AMiner 建立学者描述本体,通过拓展 FOAF 本体框架,定义包含研究者和出版物两个类型实体和 24 个属性、合作者和创作两对关系,更好地推理与挖掘学术实体间的关系,得出社交能力、活跃度等更多元特征指标[6]。为了将微软学术图谱(MAG)和 AMiner 学术图谱两个亿级异构数据进行融合,开放学术图谱(OAG)建立 venue schema、author schema、paper schema 实体和属性框架,建立 6 500 万个匹配关系,对出版者、论文和作者进行结构化数据描述[7]。

1.3 学者库构建方式

目前,大部分学者库结合自动化与众包的思想,基于数字图书馆的数据库文献资源自动构建学者库,之后采用多种方式鼓励学者人工审核与完善学者信息。

在资源组织与描述基础上,数字图书馆自动化构建学者库的关键是实现学者学术成果与学者的关联。关联过程中,不可避免地出现学者姓名歧义现象,需要区分同名学者不同的身份信息与学术资源,这也是当前研究与实践的难点。为在海量学术资源中准确定位学者及其科研成果,AMiner 采用网络分析法,根据实体关系权值,分析重名学者自我中心网络特点和属于不同团块的特性,通过集团划分来区分不同实体,实现学者人名消歧[8];中国知网、万方主要通过“姓名 + 单位”的组合方式进行学者消歧[9-10];清华大学学者库挑选具有价值的学者为其设立学者标识符 THUID,启动发文自动追踪项目,制定完整的分析和追踪策略[11];还有一些研究与实践则针对文献作者姓名的消歧方法展开探索[12-13],或是试图通过关联 ORCID、ResearcherID 等唯一标识符和建立规范文档进行学者识别[14]。

在促进学者人工审核与完善学者信息方面,当前学者库主要通过科研管理的手段和设置资源权限奖励的方法,促进学者参与科研成果注册登记。厦门大学将学者库作为科研信息管理平台,与统一身份认证平台进行数据共享,根据学者反映的信息补充、更正学者库数据[15]。ResearchGate 需要学者完成注册才能使用库内资源,通过学者自主注册与库中已有学术资源、学者信息进行匹配,提交学术成果的文档、链接或相关证明,经过审核后完成学者注册。在理论研究方面,也有研究者以机构库、学者库为基础,在学者甄别的基础上,设计学者标识、甄别匹配、推送认领、补充认领等学

者学术成果认领流程[16－17]。

1.4 学者库应用方式

大多数的学者库都设立学者检索页面和学者主页,用于展示学者的基本信息、研究成果及动态。澳门大学学者库设置 ORCID、题名、作者等 14 个检索字段,支持图片检索、高级检索和专业检索[18]。AMiner 学者库成为搜狗学术搜索数据提供者[19],增加学者数据使用频次。清华大学、兰州大学、澳门大学等高校的学者库在首页推送本机构学者在 *Cell*、*Nature* 和 *Science* 等顶尖学术期刊发表的论文;设置“推送高被引/热点文章”和“本期推荐”栏目,定期推荐热门文章和学者[20]。厦门大学学者库与科研产出相关联,成为年度绩效考核、职称评定、项目申报和管理的基础数据,设置独立评价指标库,利用可视化工具为学校管理层提供决策支持[15];清华大学学者库于 2017 年成为职称申报的学术论文数据来源和教师年终考核工作的学术论文数据源。

除了服务学者和科研部门,学者库在人才挖掘领域也得以利用。AMiner 学者库以智能服务为基础,构建国家自然科学基金委员会专家 Profile 系统,并为科技部构建专家画像库;建立阿里巴巴人才地图、CFF 专家系统,服务于企业与科研机构。ResearchGate 通过学者与机构的关联,计算机构科研水平帮助学者快速查找具有合作潜力的项目、机构与学者并提供科研招聘服务,机构与个人能够通过 ResearchGate 雇佣高质量研究人员[21]。

总之,近年来学者库得到快速发展,人名消歧、学术成果自动追踪、建立学者唯一标识符等成为学者信息及其资源采集和整合的常用技术手段和方法;人工智能、机器学习已开始运用于学者库建设与应用之中,通过语义挖掘、深度学习,建立本体或结构化的描述体系对学者进行揭示。当然学者库在建设过程中,依然存在构建方式与效果不理想的问题:①学者特征揭示不全面,重视对学术产出的集成和计量,学者学术关系的推理和学者实体特征的挖掘不深入;学者身份信息与学术资源的识别与匹配不够精确,自动追踪学者学术产出的程度不高。②数据来源单一,主要基于数字图书馆资源建立学者资源库,没有融合海量的网络资源;一些数字图书馆没能发挥出资源分布式存储与管理的优势,没有集成多方数据源全方位整合学者学术产出,无法为学者库的构建提供完整支撑。③学者库应用的推动力不足,局限于学者页面的生成、学者检索等基础功能;没有成为学者知识存档、学者轨迹展示以及机构知识库构建的有效手段;与科研管理、科研评价的结合还不够紧密,在专家发现和人才评价、绩效考核方面没有发挥出最大成效。

2 学者特征及其元数据模型

基于数字图书馆的学者库既要反映学者各项基本信息,应用于文献服务中的学者消歧,又要深刻揭示学者学术属性,为更深层次的个性化服务提供数据基础。学者库应对反映学者属性特征的元数据进行有效组织,结合应用目标和需求,从海量的学术资源中提取和识别元素,形成结构化的学者信息描述框架,需要通过对学者信息的有序组织,学术属性的识别与揭示,准确把握学者特征,动态反映学者学术轨迹。

2.1 学者特征属性分析

学者是在科学、文化、教育领域专门从事研究工作的人员[22],具有相应特征实体和属性,如接受的专业教育、拥有的高等教育学位、所在单位性质(科研院所、高校、企业研发部门等)、从事的科学研究和生产的专业领域、学科或专业特长;公开发表的论文、拥有的专利、获得学术荣誉、拥有的学术关系网络等。每个学者又因学术经历、学科领域的不同而拥有不同的特征,如人文社科类学者较少拥有发明专利。这些学者属性分散在数字图书馆学者注册信息、文献数据库、学者个人页面、学术新闻、社交网络等来源之中,可以反映学者各式各样的特征。所以,学者特征的遴选应面向数字图书馆学者库的应用需求,从学者识别、科研评价、人才挖掘、个性化服务等应用场景出发,同时注重学者学术特征的揭示的全面性,设计既能准确反映学者学术共性又能灵活反映学者个性特征的学者特征属性框架。

国内外许多研究与实证从不同角度、不同方法设计和论证学者模型或描述框架,对于学者库元数据结构和学者元素的梳理具有重要参考价值。通过对文献[23－25]和 Aminer、中国知网、百度学术、清华大学、北京大学等学者库调研,并以数据来源、学者特征和应用场景为考量因素,构建“学者维度－元素”学者特征属性框架,如图 1 所示。因为学者各属性特征出现的频次不同,构建的学者框架应该允许部分学者特征重复或缺失。使用正则表达式表达各元素出现次数规则:“*”表示 0 次或多次;“?”表示 0 次或 1 次;“+”表示 1 次或多次;无符号表示必须出现且仅 1 次。本文设计的学者库学者特征属性框架包括 6 个方面共 27 个元素:基本信息反映学者自然属性,通讯信息应用于学

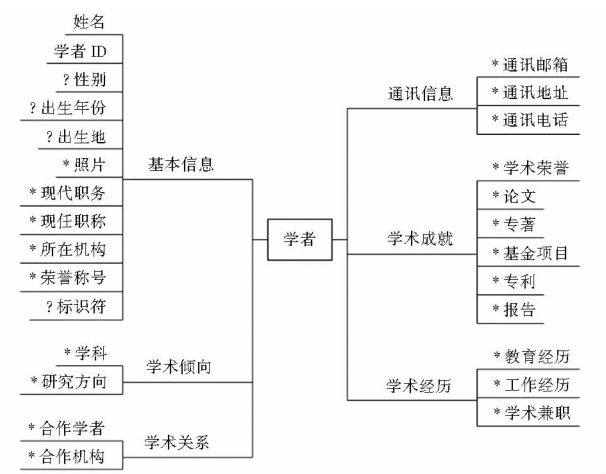

图 1 学者库学者特征属性框架

术交流、沟通和联络，二者是学者姓名规范、实现学者识别的基础数据；学术倾向反映学者研究方向、学术特长等，集成的数据可应用于数字图书馆精准科研服务；学术关系包括正式与非正式学术交流中合作的学者和机构，反映学者学术关系网和活跃度；荣誉、论文、专利、专著、基金项目等元素反映学者的学术成就，教育经历与工作经历反映学者学术背景与学术经验，是学者评价与人才挖掘的基础。

2.2 学者元数据模型

学者库的构建不仅是学者与文献数据的匹配和描述，还涉及学者、成果、机构等科研实体，不同实体与属性之间存在着逻辑关系，因此数字图书馆可以借鉴实体关系网络的方法，通过科研实体之间的链接，实现实体关系与属性的推理和挖掘。根据图 1 学者特征属性，通过实体－关系－属性的表达方式，设计如图 2 所示的数字图书馆学者元数据模型，实现学者数据的结构化表达与动态关联。将论文、荣誉等学者成果和学校、机构等单位转换为实体，并拓展每个实体的属性；学术倾向无法转换为实体，由学科和研究方向属性直接与“学者”实体进行关联；学术关系中的合作学者和合作机构可以通过论文、专利等实体中作者与机构的属性实现，一些学者属性是由学者与科研实体相结合产生的，无法归于学者或其他科研实体，应属于实体的关系。如学位、毕业时间、专业属于学者的教育经历，不是学校固有的属性；而学者对应特定学校才有相应的学位、毕业时间等属性，故这些属性应归于“学习”这个关系中。

为了实现不同来源数据的关联和存储，需要对学者元数据进行逻辑结构设计，以便构建关系型数据库。按照数据库第三范式（3NF）将学者元数据 E-R 模型转换为关系模型，且满足第一范式与第二范式，构建相关

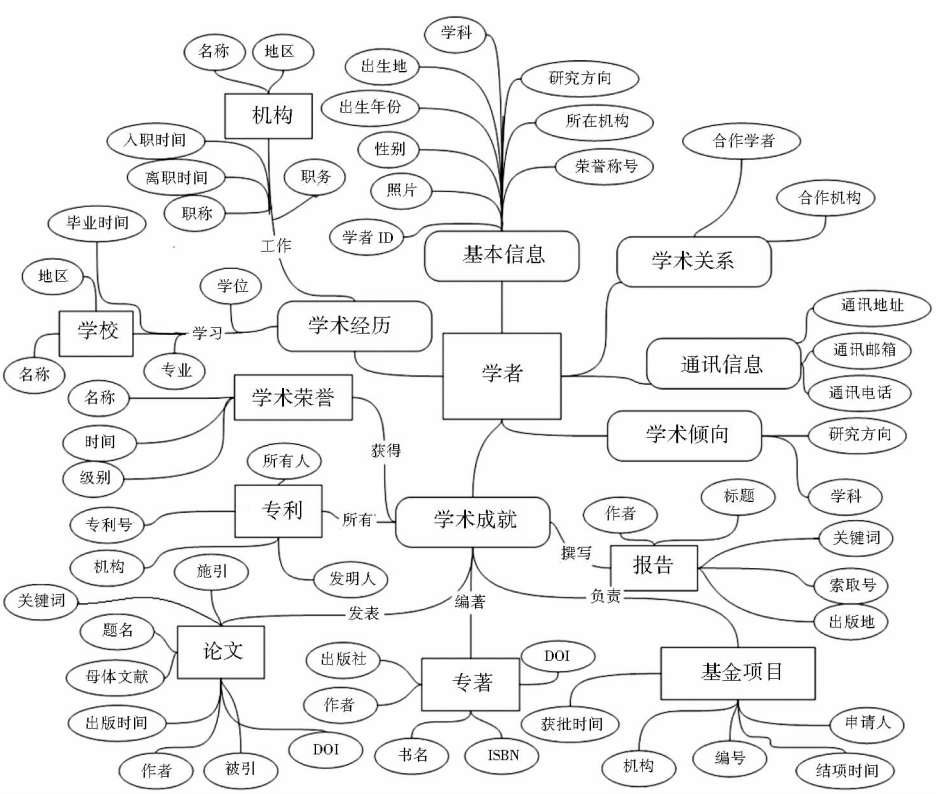

图 2 学者库学者元数据概念结构框架

数据表，实现不同数据表的关联见图3。学者ID关联学者成果信息表，可以集中快速展示学者所有成果；学者ID也是关联学者相关属性或特征的基础，这样在不同字段实现不同表间的关联，当学者数据产生更新、变动时，相关的数据表和字段进行相应更新。

图3　学者库数据关系模型

3　基于数字图书馆的学者库总体框架及构建流程

3.1　基于数字图书馆的学者库总体框架设计

基于数字图书馆构建学者库，需要依托数字图书馆的技术体系结构和信息体系结构，借助数字图书馆资源加工采集系统、异构资源整合系统、数字资源的管理系统、资源调度系统、用户管理系统等系统平台，设计学者库构建总体框架，见图4。同时汇集不同来源的学者数据，采集、加工、整合、存储学者相关学术数据、学术资源等数字对象，并进行学术网络建模分析，最终实现学者数据的应用。

数字图书馆学者库以互联网资源和数字图书馆资源为数据来源，通过数字资源采集加工系统，基于OAI-PMH协议收割学术资源元数据，收集数字化文档、出版物等数字化信息。基于数字对象系统将数字资源按照描述数字对象的条例和规则加以描述，生成元数据与调度码，共同构成数字对象。在整合层进行资源的去重合并，进行数字资源的标准化加工；借助数字图书馆异构资源整合系统，实现数字图书馆内外部元数据、资源的整合。基于数字资源管理与存储系统，根据数字图书馆分布式存储和学者学术资源多来源、多渠道分布的特点，采取元数据集中存放、数字对象分布存放的存储方式存储数据。在学者数据整合与存储的基础上对学者进行建模分析，将依据学者元数据框架进行集成，形成学者标签体系，为学者画像提供基础。以文献数据和社交网络为基础，进行挖掘与分析，从不同学者、不同学术资源间的网状关联中，构成学术网络模型，揭示合作关系网络。根据学者特征，对学者聚类，挖掘相似学者，揭示学术团队。以数字图书馆资源发布与用户检索系统为基础构建服务平台，将学者资源最终应用于学者评价、学者画像、知识管理、科研管理、学者检索和学者精准推荐等。

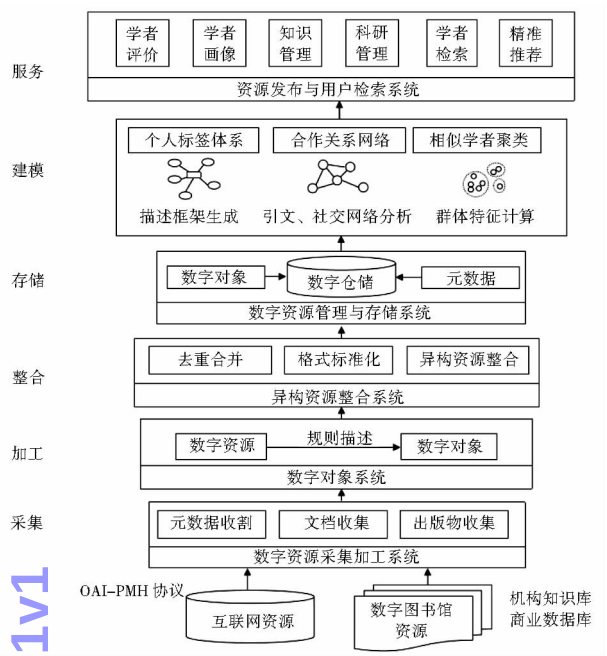

图 4 基于数字图书馆的学者库构建框架

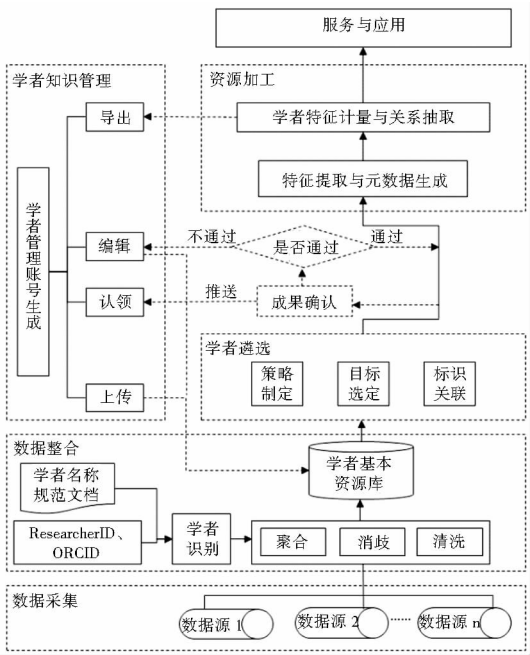

图 5 基于数字图书馆的学者库构建流程

3.2 基于数字图书馆的学者库构建流程

基于数字图书馆的学者库的构建,需依托数字图书馆自身资源与平台,对不同数据来源进行采集,通过学者名称规范文档和唯一标识符实现学者识别,对数据源进行聚合、消歧、清洗,形成学者基本资源集;在此基础上根据遴选策略选定目标学者,通过学者认领实现信息与成果的确认;通过特征挖掘和关系抽取完成对学者数据和资源的加工,最终实现学者库的服务与应用。设计数字图书馆学者库构建流程见图 5,其中,学者库构建的关键性步骤如下:

3.2.1 多源数据采集

数字图书馆应该发挥分布式资源管理的特色,与不同国内外知名数据库商合作,丰富学者库构建的基础学术文献资源;运用机器学习原理和自动追踪方式,从数字图书馆所拥有的学术文献资源中挖掘学者学术成果及利用信息;同时,发现和收集网络资源中学者主页、人物百科、学术新闻等学者相关网页,丰富和完善学者相关信息,获取其最新的学术动态。学者库建设不是一蹴而就的,需要建立信息采集的更新机制,持续进行资源的采集与更新;依据互联网页面的布局及对应的学者元数据变化,建立信息抓取监测机制,及时完善数据抓取中的问题。

3.2.2 学者数据整合

对采集的多来源学术信息数据进行清洗、整合与基于学者的聚合,是学者库资源建设环节的重要工作。数据清洗环节的主要任务是实现采集数据的规范化,剔除低质量的数据,补充缺失字段等。数据整合环节的主要任务则是将不同来源的数据汇聚,其重复数据对于存在部分字段不一致的数据进行冲突处理。在此基础上,借助 ORCID、ResearcherID 等学者唯一标识符,学者名称规范文档以及机器学习技术进行学者姓名消歧、资源与学者的关联,以实现学者库资源的精准、全面聚合。

3.2.3 学者遴选

以数字图书馆收录文献的作者作为遴选范围,针对学者库建设的不同应用目标,需要设置发文、被引阈值等定量指标或学者身份等定性指标,制定学者遴选标准。从学者身份、学术成就、专业技能等角度设计学者遴选策略,挑选出有收录价值学者,如高产、高被引作者或院士、“千人计划”、学科带头人等热门学者作为重点收录目标。使用标识符关联目标学者,可以根据需求对在库学者进行编码,或者直接与 ORCID、ResearcherID 等常用唯一标识符进行链接,对遴选学者进行动态更新,更新遴选对象与范围。对文献元数据和学者元数据的提取与加工,突出了学者特征,形成可读取、可储存、可关联、可展示的学者元数据。

3.2.4 特征挖掘与关系抽取

以遴选学者为基础,参照建立的学者数据逻辑结构框架,使用命名实体识别技术识别学者的相关学术实体、属性及关系,并进行实体抽取与属性抽取。根据学者元数据进行统计与推理,挖掘学者的学术属性特征。对学者个人身份特征进行梳理,对学术情况进行计量,对学者间的特征信息进行关系计算,不仅形成如

发文量、h 指数等学术评价指标和工作经历等学者学术线性的发展轨迹，还可形成合作、引用等网状的学术关系。继而进行学者间引文关系、合作关系、社交关系的挖掘与分析，抽取学者与各科研实体间的学术关系，建立学术关系网络模型。

3.2.5 成果认领与管理

在学者相关数据集成后，需要对整合后的学者信息进行确认。但图书馆无法强制学者使用学者库，对整合后的成果进行确认、对个人信息进行维护，故该环节只针对使用学者库的学者或联合科研管理部门进行。需要通过政策激励、服务升级，引导和吸引学者完成成果认领。引导学者通过学者库知识管理平台，完成学术成果的认领、个人信息编辑和修改以及学术成果的统计与导出。采取机器学习与人工审核相结合的方式对 学者学术成果进行验证。对于注册加入学者库的学者，将整合后的学者信息推送至学者账号，学者对资源进行认领。若审核通过，则对学者信息进行特征提取；若不通过，则允许学者对进行编辑，并重新整合至学者信息集合中，实现循环的审核与更新机制。

4 基于数字图书馆的学者库构建与应用推进策略研究

学者库的建设与应用是相辅相成、循环渐进的动态过程，应按照“边建设、边使用、边完善”原则进行学者库构建与应用的同步推进。为改善学者库效果，需要激励学者积极参与信息的完善与审核；对接科研管理平台，提高学者库构建基础数据的质量。面向管理机构，可以推进其学者库在科研过程管理、人才管理、资源建设中的应用；面向学者，可以推进学者库在其学者知识管理、学术信息资源精准服务中的应用。

4.1 增加展示与宣传效果，增强学者参与动力

受入库资源质量及技术限制，全面准确地采集学者信息、进行高精度的学者姓名消歧仍是难点，因此学者库需要提升学者建设与使用学者库的参与度，才能提升学者数据构建的全面性和准确性。学者页面与个人的学术形象息息相关，能够吸引学者丰富和维护个人的信息，从而提升数据准确性。将学者及其信息的展示作为增强学者参与学者库构建与应用的动力，在学者页面通过计量分析、可视化展示等手段，帮助学者提升学术影响力；推送热门学者主页，增加对学者库个人展示功能的宣传，激发学者成果认领、信息维护完善个人主页的热情；吸引学者使用学者库资源而产生的访问、浏览、下载等行为数据可以作为资源质量评价的参考。通过学者的认领、应用和互动，提升学者库信息质量。

4.2 搭建科研管理平台，融入科研管理过程

与科研管理结合，既可以服务科研管理部门，也有助于丰富和完善学者库信息，提升学者库质量。将学者库构建融入成果收集、成果考核、科研评价、项目申报等科研管理环节，作为学术成果提交和职称评定和科研考核、项目申请的基础数据，方便和优化机构内部科研绩效管理，进行学者学术产出统计与管理。同时对学者填报信息逐一审查，确保学者信息和学术资源的完整性和准确性，形成科研信息申报审查机制，可以强化学术规范，避免科研失信。此外，科研管理平台中的信息都是学者确认后的、时效性较强的信息，因此可以将其作为学者库构建的数据来源，提升入库信息质量。

4.3 与人才识别相结合，服务于团队和专题资源建设

学者库对学者进行特征挖掘、关系抽取，按领域、学科、专业或单位对学者进行有效类分，可以识别和发现学者擅长、精通和潜在学术领域，应用于不同学科的人才识别与人才选择，成为专家遴选、科研评审、项目支持的专家人才储备库。针对机构学者进行资源建设，集成某一机构或某一领域的专业学者，形成“专、精、深”的学科专题资源库，拓展机构知识库特色资源。

4.4 与知识管理相结合，兼顾学者存档功能

对学者而言，学者库囊括了学者自身的相关学术信息和成果，是学者知识管理的工具和平台，也是学者有效存储个人知识的场所，可以作为开放获取自存储实现的绿色仓储；实现学者学术成果的添加、编辑、删除，将学者库打造成学者个人知识库；不仅将学者的学术资源进行集成，还对学者信息进行结构化梳理，帮助学者厘清学术发展路线。

4.5 构建学者画像与学者模型，拓展精准服务功能

将学者库嵌入知识发现、科研管理、学术社区等科研创新平台，能更好地为学者和科研机构提供信息服务。学者库集成不同来源的学者资源，进行学者识别，实现学术资源的姓名消歧，能提供学者及其成果的搜索和发现服务；以学者为单位组织资源，从不同角度刻画学者学术特征，能够为科研管理平台提供基础数据，提供学者计量和评价服务；运用学者库数据挖掘学者学科兴趣、发展趋势等，构建学者画像和用户信息模型，逼近学者客观实际，为数字图书馆学者精准资源推送服务奠定基础，推荐相关学者，促进学者交流与合作。

参考文献：

[1] 曾建勋. 加强学者库的建设与应用[J]. 数字图书馆论坛,2018,(9):1－1.

[2] CNKI. 成果库帮助[EB/OL]. [2019－04－21]. http://papers.cnki.net/Help.aspx.

[3] 西安交通大学图书馆. 机构知识门户开放获取政策[EB/OL]. [2019-08-27]. http://www.ir.xjtu.edu.cn/web/policies.
[4] 窦天芳, 张成昱, 张蓓. ResearcherID 现状分析及应用启发[J]. 图书情报工作, 2014, 58(4):40-45.
[5] 刘敏,田丽. 开放学术环境下高校学者库建设路径研究[J]. 图书馆学研究,2018(20):28-34.
[6] TANG J, ZHANG D, YAO L. Social network extraction of academic researchers[C]//IEEE international conference on data mining (ICDM 2007). Omaha, NE:IEEE,2007. 292-301.
[7] Open academic graph | Open academic society. [EB/OL]. [2019-04-15]. https://www.openacademic.ai/oag/.
[8] TANG J, ZHANG J, YAO L, et al. ArnetMiner: extraction and mining of academic social networks[C]// Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas, Nevada: ACM, 2008. 990-99.
[9] 成果库帮助[EB/OL]. [2019-04-21]. http://papers.cnki.net/Help.aspx.
[10] 认领成果[EB/OL]. [2019-04-21]. http://social.wanfangdata.com.cn/index/toFistBroInfo.do.
[11] ThuRID(Tsinghua University Researcher ID)服务目标[EB/OL]. [2019-04-15]. http://rid.lib.tsinghua.edu.cn/static/web/xls_template/constructionGoals.pdf.
[12] 肖晶, 梁冰, 张晓丹. 一种面向篇级数据的作者名消歧规则和算法[J]. 数据分析与知识发现, 2012, 28(5):55-59.
[13] 郑威杰. 科技文献作者消歧方法研究[D]. 杭州:杭州电子科技大学,2017.
[14] 常娥. 学者身份识别的机制及关键技术研究[J]. 图书馆论坛, 2015(10):88-95.
[15] 杨薇, 林静, 黄国凡,等. 面向"双一流"建设的学科知识服务营销策略——厦门大学图书馆的实践[J]. 大学图书馆学报, 2017, 35(5):74-79.
[16] 张雪蕾, 魏青山, 陈雅迪. 基于甄别算法的学者学术成果认领机制的研究与实践[J]. 情报理论与实践, 2018,41(02):68-72.
[17] 刘巍, 祝忠明, 张旺强, 等. 机构知识库中作者标识与作品认领机制的研究与实现[J]. 现代图书情报技术, 2014(3):8-13.
[18] 澳门大学学者库[EB/OL]. [2019-02-15]. http://repository.umac.mo/advanced-search.
[19] 申明. AMiner:科研搜索"神器"[N]. 科技日报, 2018-06-12(6).
[20] 兰州大学学者库. [EB/OL]. [2019-02-15]. http://202.201.7.4/browse-author.
[21] Scientific recruitment- hire researchers & scientists on researchgate. [EB/OL]. [2019-02-15]. https://www.researchgate.net/scientific-recruitment.
[22] 李旋. 军事医学高学术影响力学者的识别方法与揭示架构研究[D]. 北京:中国人民解放军军事医学科学院,2017.
[23] 李纲, 叶光辉. 多源专家特征信息融合研究[J]. 数据分析与知识发现, 2014, 30(4):27-33.
[24] 宋培彦, 陈白雪, 贤信. 科技专家信息语义模型构建及实证研究[J]. 情报理论与实践, 2017, 40(9):119-124.
[25] 范晓玉, 窦永香, 赵捧末, 等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作, 2018,62(15):31-40.

作者贡献说明:

郑昂:负责选题调研与分析,撰写论文;
曾建勋:提出论文选题与思路,修改论文。

Study on the Construction Method of Scholars Repository Based on Digital Library

Zheng Ang Zeng Jianxun
Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] From the perspective of the utilization and arrangement of digital library resources, this paper designs the construction mode of scholar data identification and scholar repository, which helps to improve the efficiency and characteristic service of digital library resources construction. [**Method/process**] This paper investigated the research and practice of scholar database at home and abroad from 4 aspects: the source of scholar selection and collection, the content and framework of scholar description, the construction and application of scholar repository. By analyzing the characteristics and attributes of scholars, studying the structured expression of scholars´data, this paper put forward the construction process and overall framework of scholar repository based on digital library. [**Result/conclusion**] This paper puts forward the promotion strategy of building and application of scholar repository, emphasizes that scholar repository should be integrated into the scientific research management process, mobilize scholars to participate in the construction, increase the effect of exhibition and publicity, combine with talent identification, serve the construction of team and thematic resources. It is also important to combine with knowledge management, take into account the functions of scholar archive and scholar portrait, and expand the precise service function.

Keywords: scholars repository digital library institutional repository scholar identification